RESEARCH ARTICLE

'Everything Counts in Large Amounts': A Case Study of the Mechanisms of Data-based Production

Aleksi Aaltonen
The London School of Economics and Political Science
aake@iki.fi

Niccoló Tempini
The London School of Economics and Political Science
n.tempini@lse.ac.uk

Corresponding author

Aleksi Aaltonen
The London School of Economics and Political Science
Houghton Street, London WC2A 2AE, UK

aake@iki.fi
+44 796 550 2511

Suggested running title
Mechanisms of Data-based Production

Acknowledgements
The authors would like to thank Jannis Kallinikos and Carsten Sørensen for their support and feedback.

This is a post-peer-review, pre-copyedit version of an article published in *Journal of Information Technology*. The definitive publisher-authenticated version Aaltonen, A. and Tempini, N. (2014). Everything Counts in Large Amounts: A Critical Realist Case Study on Data-based Production. *Journal of Information Technology*, 29(1): 97–110. is available online at: http://dx.doi.org/10.1057/jit.2013.29

‘Everything Counts in Large Amounts’: A Case Study of the Mechanisms of Data-based Production

Abstract (250 words)

Contemporary digital ecosystems produce vast amounts of data every day. The data are often no more than microscopic log entries generated by the elements of an information infrastructure or system. While such records may represent a variety of things outside the system, their powers go beyond the capacity to carry semantic content. In this article, we harness critical realism to explain how such data comes to matter in specific business operations. We analyse the production of an advertising audience from data tokens extracted from a telecommunications network. The research is based on an intensive case study of a mobile network operator that tries to turn its subscribers into an advertising audience. We identify three mechanisms that shape data-based production and three properties that characterize the underlying pool of data. The findings advance the understanding of many organizational settings that are centred on data processing.

Keywords (up to six)

Audience, Critical realism, Data-driven, Information actualization, Measurement, Mechanism

INTRODUCTION

Prominent IS scholars have repeatedly complained about weak theoretical foundations for analysing the mutual constitution of technological systems, organizational arrangements and outputs (e.g. Lyytinen and Yoo, 2002; Orlikowski and Barley, 2001; Yoo, 2010). In order to cope with the problem, researchers continue to import theories from other disciplines, whereas attempts to strengthen theory building capacity within IS are rarer (Baskerville and Myers, 2002; Benbasat

and Zmud, 2003; Lee, 2010). In the spirit of the latter approach, this article demonstrates how critical realism (CR) helps to build a theoretical explanation of a specific, data-driven product innovation in commercial media. CR works as a metatheory[i] for our study. It is not concerned with specific empirical phenomena but is rather a theory of ontology and epistemology that guides the construction of theoretical explanations. Critical realism provides a robust, explicit framework for theorizing causal mechanisms that underpin a new kind of advertising audience.

The analysis revolves around a start-up telecommunications operator that has built a new form of commercial media by relaying advertisements to mobile phones as text and picture messages. The challenge for the company is that sending marketing messages to consumers does not yet constitute a viable medium for advertising. This is because advertisers are not willing to pay for advertising to an unknown audience (Ettema and Whitney, 1994; Napoli, 2003). Any aspiring media company must know its audience along relevant dimensions – otherwise it cannot sell media space to advertisers. This knowledge is typically based on a sophisticated technological capacity to monitor people's exposure to media content and advertisements. The opportunity for the company to construct an audience is grounded on its access to data from a telecommunications network infrastructure. To understand the emergence of a new kind of advertising audience, we ask the question:

> *What mechanisms allow the company to manufacture an advertising audience from the mobile network data?*

The idea of audience is a slippery concept that has no single accepted definition (Bratich 2005; Morley 2006). In this article, we understand an audience first and foremost as a product. The business of media companies is about creating, maintaining and selling audiences to advertisers. This is made possible by audience measurement arrangements, whose evolution has historically shaped media products, content and the whole industry (Bermejo, 2009; Carr 2008; Napoli 2003, p.

83). For this purpose, a mobile network infrastructure has the special feature of generating data tokens known as call detail records (CDRs); these capture network subscriber behaviour in a microscopic, standardized way across network elements. Yet, as we will show below, CDRs are meaningless in the context of organizational practices. No relevant pattern or insight emerges by looking at the raw data tokens. In order to have a product to sell for the advertisers, the company must turn the data into information about an audience.

The data tokens can be understood as non-material technological objects (Faulkner and Runde, 2009, 2010, 2013; Runde, Jones, Munir and Nikolychuk, 2009) or digital objects (Ekbia, 2009; Kallinikos, Aaltonen and Marton, 2013). The concept of object is central to critical realist theorizing and connects the study with recent discussions on materiality (Leonardi, Nardi and Kallinikos, 2012; Mutch, 2010; Orlikowski, 2007). We assume that the data tokens have syntactic properties that make a concrete impact on the audiencemaking operations. These properties neither derive from the physical medium storing the data nor are simply representations of external reality. Indeed, we argue that the data are 'material' in the adjectival sense that they matter beyond their semantic content. Phenomena like those that we set out to investigate are the focus of what has also been called digital materiality (Yoo, Boland, Lyytinen and Majchrzak, 2012). Advertising audiences certainly have a lot to do with people using media content, but the variables that ultimately construct the audience product in the industry have always been influenced by technological measurement arrangements (Ettema and Whitney, 1994).

The article makes two contributions. First, we show how the critical realist framework supports the theorizing of causal mechanisms which are activated in the audiencemaking process (Bhaskar, 2008; Sayer, 2000). The term ‘audiencemaking’ is used throughout the article as shorthand for the construction of an audience as a product (Ettema and Whitney, 1994). Second, the properties of digital data and related causal mechanisms that emerge from them are not idiosyncratic to this case study. Given the relatively generic nature of data tokens such as log entries across

different systems, our results can inform studies focusing on a wide variety of settings.

CRITICAL REALISM

Critical realism (CR) is a philosophy of science that has a set of basic principles at its core (Archer, 1998; Bhaskar, 2008; Mingers, 2004; Sayer, 2000). The approach makes two fundamental assumptions with respect to the methodology of empirical research: first, the world exists independently from our knowledge; second, the world can be observed only partially. CR can be thus seen as drawing from the constructivist critique to earlier forms of realism, holding that both researchers and their informants encounter the world through interpretation (Sismondo, 1993, p. 535). Importantly, however, CR also holds that those interpretations can carry traces of a reality that is independent of present actors. This allows CR to incorporate the idea that all knowledge is socially constructed and thus *transitive*, while scientific knowledge addresses *intransitive* structures of reality that do not depend on individual awareness of them and are independent from any given context. The difference between transitive knowledge and intransitive reality is central to CR and will be discussed below.

Transitive knowledge about intransitive reality

There would be little point in CR if the intransitive reality simply mapped to natural phenomena while all artificial (Simon, 1996) were considered transitive. Quite the contrary, the intransitive reality is very much populated by the outcomes of human actions and interpretations. Let us call these relatively stable human-made entities 'social structures'.

All action depends on structures. Archer (1998, p. 197) points to Bhaskar, who "states unambiguously that 'social forms are a necessary condition for any intentional act, (and) that their pre-existence establishes their autonomy as possible objects of scientific investigation'". Social structures enable and shape actions, and as such they are important objects of scientific research. Entities such as a cultural

convention, technological infrastructure or a law can have a structuring effect on action. Structures originate and are reproduced in human activities. Nevertheless, CR differs from popular IS approaches, such as structuration theory, actor-network theory and sociomaterial perspectives, in that it rejects the conflation of structure and action. An action cannot draw upon a structure and simultaneously bring it into existence (Archer, 1982; Mutch, 2010).

The separation of action from structure is described in the transformational model of social activity (TMSA). The model describes how action draws upon, reproduces and changes structures in a temporal sequence (Faulkner and Runde, 2013; Runde, Jones, Munir and Nikolychuk, 2009). In our analysis, the focus is on the implications of an already existing structure (CDR data) on audiencemaking. We are interested in understanding mechanisms that emerge from the structure in a particular setting rather than in structural transformation. Consequently, we demarcate the case so that the construction of the CDR infrastructure is excluded from the analysis. This is also justified by the fact the infrastructure is taken as a given for all practical purposes at the research site. The CDRs are, in the language of TMSA, a structural condition for the company operations.

In the critical realist framework, CDR data objects, the instantiation of audiencemaking events and empirical observations map to different epistemological domains. The approach postulates an ontology in which the phenomena of scientific interest are structured beyond their empirical appearances. Obviously, many things that exist can be observed, but the existence of something does not depend on its observability. The most fundamental structures and mechanisms can often be established only analytically (Bhaskar, 1998, p. 41; Mingers, 2004, p. 93). According to CR, the intransitive reality – reality which is distinguished from the scientific discourses around it – is stratified into the *real*, *actual* and *empirical* domains. These are nested so that the real contains the actual, which contains the empirical. The domains allow different epistemic access, which has profound methodological implications. The empirical domain can be accessed by direct observation, while the

actual and real domains are investigated through retroductive theorizing that we will introduce below. The purpose of research is usually to uncover structures and mechanisms that account for relevant events, some of which are captured in empirical observations.

Stratified ontology

The domain of the real consists of objects, and mechanisms that arise from them. A structure is constituted by a group of component objects, which are interrelated in a specific configuration. A structure is an object itself because it expresses *emergent* properties that cannot be reduced to the individual components of the structure (Elder-Vass, 2007). For instance, an organization is a structure that can have the capacity of producing aeroplanes, while none of its individual units or members has such a capacity alone. Component objects, such as organizational units in the example, are often internally structured in their turn. The constitutive associations that make an object/structure are called internal relations, whereas objects often have many external relations that do not affect their constitution or properties (Easton, 2010; Faulker and Runde 2013; Wynn and Williams, 2012). A collection of objects that expresses only the *resultant* properties of its parts is not a structure but an unstructured aggregate (Elder-Vass, 2005). Structures sustain mechanisms that account for causality and are the primary interest of scientific explanation. A mechanism can be understood as a capacity, that is to say, a possibility or tendency of what is likely to happen under certain conditions (Wynn and Williams, 2012, p. 791). Mechanisms are causal powers and must be activated for certain events to happen. Moreover, since objects/structures are continuants, they can sustain causal powers across time and space (Easton, 2010; see also Faulkner and Runde, 2010).

In order to illustrate these abstract concepts, let us make some preliminary distinctions in the arrangements underpinning audiencemaking operations at the research site. To begin with, the telecommunications network infrastructure routinely generates a massive amount of individual CDR data tokens. These can be understood as relatively simple objects. Together, the CDRs constitute a data mass

that may express emergent properties. The data are hence a potential structure, which can give rise to mechanisms that are relevant in audiencemaking. We call this candidate structure a 'data pool'. Our intention is then to investigate if the data pool has emergent properties that give rise to mechanisms shaping audiencemaking events and, ultimately, the audience product.

Events stem from the activation of mechanisms. It is worth emphasizing that the concept of event in CR is broad. For instance, "a bad year, a merger, a decision, a meeting, a conversation, or a handshake" can constitute an event that requires an explanation (Langley, 1999, p. 693; see also Wynn and Williams, 2012, p. 786). An event may happen only once or may be representative of a series of events that stem from the same mechanism. The kind of event to be explained depends on the research question that a study addresses. The domain of the *actual* contains all the events that take place, both those that are observable and those that remain unobserved, whereas the empirical domain covers only the events that are observable. The latter provide the starting point for critical realist theorizing about underlying structures and mechanisms.

Retroductive reasoning

Retroductive reasoning starts from an observed event and moves to theorizing the "hypothetical mechanisms that, if they existed, would generate or cause that which is to be explained" (Mingers, 2004, p. 94–95). The cause of an event is considered to be what makes a difference to its realization. However, it is important to note that causal explanations are usually focused only on certain mechanics behind the event (Runde, 1998). It is often more interesting to analyse the event for specific features rather than whether or not it happens, or to try listing every possible mechanism involved. For instance, a press release is an event that is shaped by such factors as linguistic structures, public relations practices, managerial authority and a particular distribution channel. Yet, in this research we are interested in press releases and other events for the ways in which they contribute to the construction of an audience product. The same event can be accounted for in many different

arguments, each focusing on a different aspect of the event and consequently providing a different kind of explanation.

Retroductive reasoning starts from empirical observations of an event. It then proceeds by analytically reconstructing mechanisms that would explain the event. The resulting explanation does not have to exhaust all aspects of the event, but it must be expressed in a way that allows the testing of its validity through further empirical studies. Theoretical explanations can compete when they result from attempts to capture the same structure or mechanism from different angles (Sayer 2000, p. 11), and they may eventually explain aspects of the structure that other theories ignore. However, the possibility of multiple theoretical explanations does not mean their equivalence. CR rejects a strong relativist position; its epistemic relativism does not imply judgmental relativism (Mingers, 2004). Competing explanations can and should be compared, for the most accurate account of relevant causal mechanisms should have the highest explanatory power (Runde, 1998).

What makes discovery and validation difficult is that an activated mechanism may produce events that do not become observable in the empirical domain. There are often countervailing mechanisms that counteract or impede the manifestation of a mechanism to the observer. The regular observability of an event generated by a causal mechanism should therefore be considered a special case and not a prerequisite for a causal explanation (Runde, 1998, p. 153). The assessment of rival explanations should not depend on event regularities. Instead, a causal explanation must undergo a validation process that evaluates it according to different philosophical principles.

EMPIRICAL ANALYSIS

Our research site is a telecommunications operator that tries to turn its network subscribers into an advertising audience, that is, a product that can be sold to advertisers. The company was incorporated in 2006 after raising millions of euros in venture capital to launch a new kind of advertising platform. Operating as a

mobile virtual network operator [ii] but making money from advertising, the organization has "the soul of commercial media, but the body and muscles of a telecoms operator", as one of the informants phrased it. Consumers could sign up for the service by providing a simple demographic profile and opting-in to receive advertisements on their mobile phones, while the company offered free voice call minutes and text messages in exchange.

Research design and empirical evidence

Case study makes it possible to examine phenomena in their complexity, without reducing the object of research into just a few variables (Yin, 2003). This is an important advantage and makes the methodology compatible with a critical realist metatheory. CR supports intensive research that aims to identify and elaborate causal mechanisms rather than to quantify their efficacy (Easton, 2010; Wynn and Williams, 2012). Critical realist case studies typically answer *how* and *why* types of questions. They are suitable for unpacking circumstances in which the number of potentially relevant factors cannot be *a priori* narrowed down. An intensive case study like ours does not require a rigid explanatory framework to be fixed in advance, as its purpose is often to identify new explanatory mechanisms hidden from existing theories (Sayer, 2000).

The data collection took place during three-months' fieldwork using a variety of methods. One of the authors attended during regular working hours at the company headquarters, where he could constantly observe the 28 employees and directors located at the site. The staff consisted of experienced professionals in the fields of telecommunications, digital marketing, public relations, software development, business law, finance and management, organized into six teams responsible for different organizational functions. An observation log was constantly open on the observer's computer, allowing him to transcribe episodes as they unfolded and to avoid relying on his recollection after office hours. We define an episode as an uninterrupted sequence of interactions that revolve around a common topic. Many

(but not all) of the observed episodes can be understood as events that contributed to the effort to maintain a viable audience product.

At the beginning of the observation period, we had a broad interest in technology and business model innovation at the intersection of telecommunications and media industries. We quickly became sensitive to the role of audience measurement and, consequently, we narrowed down our focus to audiencemaking practices. These often drew on various measurement operations, tools and data. The observations were coded after the fieldwork period using a coding scheme derived from provisional explanatory ideas that emerged during the fieldwork. The purpose of the coding was instrumental rather than analytical. It allowed easy access to the episodes and gave proportions to the evidence, but the content and relationships between the codes are not central to the analysis. The process resulted in 689 episodes over 62 days of observation.

We interviewed 26 out of 28 people working at the research site; some informants were interviewed twice. The semi-structured interviews lasted from half to one hour and were based on a topical guide adjusted for each informant. The sessions were similar in structure, but the questions were tailored to the different roles covered by the informants and were designed to capitalize on recent developments at the research site. In order to map major events in the short corporate history and to achieve an insight into how the organization presented itself to advertisers, we stored all the press releases and blog posts published on the company website. The observer also exploited serendipitous opportunities for gathering additional material. He stored documents and web pages, photographed events at the office, took screenshots from information systems, and asked employees to provide examples of their instant messaging logs. Finally, we steered the fieldwork process on the basis of preliminary analysis. Every Sunday, the observer wrote an analytical memo (Walsh, 1998) reflecting upon the past week's efforts, identifying any problems or insights that should be addressed the following week. The summary of empirical evidence is presented in Table 1.

<Insert Table 1 here>

In contrast to relatively clear methodological principles on how theories can be used as explanatory devices, refined and rejected, procedures for theory building are generally less formalized (Weick, 1995). Critical realism is particularly supportive in this respect, for it offers clear principles on how to theorize substantive phenomena (Bygstad, 2010; Easton, 2010; Wynn and Williams, 2012). The process starts with the identification and explanation of events which would contribute to answering the research question, and then moves to describing mechanisms and structures that are expected to underpin those events. The former represent that which is to be explained (*explanandum*), while the latter provide the footing on which the explanation is built. A central part of critical realist analysis is retroductive reasoning, which moves from observations of events to hypotheses about mechanisms that could account for them. Finally, the hypothesized mechanisms need to be validated. Many critical realist scholars insist that the validation process should start within the study, but ultimately theoretical explanations need to be corroborated by other researchers and their independent investigations.

We conceive the retroductive identification of mechanisms as a process in which the researcher imaginatively fills the gaps between observed events with a causal account. The account explains what mechanism would produce the observed events and what structure would activate such a mechanism. For this purpose, we write an analytical narrative as a form of retroductive reasoning (Becker, 2007; Brewer, 2000). The narrative provides a medium in which it is possible to bring distinct observations together into an account informed by the critical realist metatheory. We start from specific audiencemaking events and reconstruct their connections with measurement data, gradually carving out three mechanisms operating at the research site. The weekly analytical memos made it possible for the process to be started already during the fieldwork. We allowed the past week's observations to inspire reflection and tentative explanations, which motivated attempts to fill gaps

in provisional explanations during the following weeks. The resulting account is constructed to make relevant, empirically observed events intelligible by reconstructing their underlying causal mechanisms. The analytical rigor of the narrative is safeguarded by triangulation and two further guidelines. The variety of empirical evidence allowed us to triangulate observations and therefore build confidence in our identification of important events and their features (Flick, 2004; Wynn and Williams, 2012). We also devised two guidelines to steer retroductive reasoning through our case. The guidelines helped to bring empirical evidence together systematically and to explore the meaning of most relevant tasks, operations and practices, while ignoring many fascinating but disparate episodes.

The first guideline is that the analysis should focus on events that are essential in terms of organizational survival. The viability of the enterprise would be decided by its success in attracting consumers and selling their attention to advertisers, that is, the execution of its novel business model. While the fieldwork deeply embedded us in the local setting and its shifting priorities, we identify relevant events as those that are necessary to sustain key business processes in the industrial context in which the enterprise operates. We call these *audiencemaking* events. Focusing on such events at the expense of others is consistent with the idea that retroductive reasoning does not have to account for all the structures and mechanisms present at the research site (Runde, 1998). The second guideline draws from the nature of the media industry and assumes that the importance of audience measurement has not vanished despite changes that are happening in the industry (Bermejo, 2009; Carr, 2008). The measurement of media consumption remains a central part of any effort to create a new kind of audience product. This further narrows our focus to the traces of measurement and analytical operations in audiencemaking events.

Audiencemaking events

Let us start from a mundane episode that reveals a common feature in many work practices at the research site. The audience, either as a generic ‘audience member’ or as aggregate ‘members’, is referred to, called upon and related with in daily

operations. Such episodes occur frequently throughout the day and can be readily reported from the collected empirical evidence. The episodes designate events in which the new kind of audience is articulated along various dimensions. The audience does not come into being in a singular momentous event, but in a series of small episodes by which it is incrementally reinforced and shaped. For instance, in the following episode an employee (MCM) describes technological arrangements that are used to monitor the network subscribers (informants are represented by acronyms in the excerpts).

> MCM discusses different member reporting models. At the moment there are three levels: ad hoc [manual], using dedicated reporting software and fully automatic. He talks also about the profiling of members for different countries. MCM says that a traditional operator does not care if the subscriber is away from the network for a few weeks, if the phone settings are correct, or if the phone model is up to date or not. While the operator may lose some revenue, it does not incur any costs. Therefore, it does not try to activate the subscriber. For us the consumers are the audience, for which we should have the connection.
>
> (Observation log, 16:15 on 24 March 2009)

The excerpt shows how talk between employees routinely constructs network subscribers as members. We triangulated this observation between different kinds of episodes and documents, which confirmed that 'members' are discussed across the teams as well as in external communications. They represent the basic unit of the audience, and hence we call the instantiation of an audience member in organizational processes an audiencemaking event. The audience acquires its dimensions, is targeted with interventions and justified for various purposes by such events; in other words, the audience exists by virtue of continuous production of audiencemaking events. People who subscribe to the mobile network are (obviously) never physically present, and it is from the information about their behaviour, rather than the human beings *per se*, that the audience is manufactured. The events include all kinds of interactions, operations and communications that

occur in the company, from casual discussions and whiteboard scribblings to PowerPoint presentations, Excel spreadsheets and the release of marketing materials.

One might object that the audience is best understood as an interpretive construct in the context of organizational practices. However, this is simply not how members are experienced at the research site. The audience often react unpredictably to advertising and other corporate interventions. Some advertisements are even intended to build dialogue based on members' previous answers. Others get unsolicited responses. Feedback mechanisms are so common that audience reactions are regularly factored *a priori* into plans; the employees treat the member as an interactive entity, anticipating unexpected reactions. This can be observed in the ways in which employees harness a variety of reporting tools to get their work done. We identified 11 different systems for analysing and reporting from various sources of data. These include systems to track the delivery of advertising messages and member activity, to log and follow up the resolution of network issues and generic work orders, to create software development items and test cases, to measure the usage of company websites, or to monitor the company's reputation on the web. But, as we now proceed to argue, these tools would be of little support without the constant flow of fresh data.

Data token object

A digital telecommunications network makes a record of every click, call and message relayed through it, generating millions of records every day. These are known as call details records (CDRs). A network infrastructure needs to log traffic for various purposes, such as allowing the optimal allocation of resources, detecting and recovering from malfunctions, and identifying potentially harmful activity. The existence of such records is thus a structural pre-condition related to the functioning of the network infrastructure, rather than a decision by the company that harnesses the data to enable business model innovation. Therefore, while the records make the new kind of media business practicable, the genesis of CDR

production falls outside the scope of the current investigation. The example below (taken from an unrelated specification document) illustrates the type of behavioural data that is generated by the telecommunications infrastructure[iii].

> 097369D2D7372762D310800000000000000001;1;33668741168;3322208;6;20081101004923;20081101004923;20081101004923
>
> (CDR data token generated by a digital network infrastructure[iv])

The record captures the time, type, the sending and receiving ends of a network interaction, and a few technical details about the operation. The data token carries no reference to the social settings, intentions and activities that triggered the events that are captured in the data. Indeed, a CDR data token is a sort of receipt. It represents the delivery of an advertisement, or a network subscriber's response to it, as a text message. CDRs set the digital network infrastructure apart from traditional audience measurement arrangements in two ways. First, broadcasting advertising audiences used to be constructed from measurements of the reception of programme content, which can only indirectly reveal potential exposure to advertising that takes place during commercial breaks. Second, CDRs do not just measure exposure, but they also verify the individual responses to a specific advertisement.

The data is also extremely granular with respect to any practical purpose; CDRs merely turn ephemeral behavioural events into strings of alphanumeric characters that carry little meaningful content as such. The production of audience measurement data happens at this microscopic level of digital transmission receipts. The data record behaviour at a considerably higher resolution than previous audience measurement arrangements, well below the level of individual audience members. The raw data leaves open a massive gap between the tokens and a coherent audience product. Individual CDRs have none of the rich meanings the audience and its members carry in the context of organizational practices. A single reply to an advertising message, as captured by a data token, tells nothing

organizationally relevant until it is combined with many others and is embedded into the context of a particular advertisement, campaign and a target group.

Data-driven mechanisms in audiencemaking

Next, we analyse several audiencemaking events and identify three mechanisms that enable an advertising audience to emerge from the data. The analysis builds toward a causal explanation of how advertising audiences are manufactured in the digital ecosystem. The identification and elaboration of mechanisms is also of key importance in demonstrating whether the data pool is merely an aggregate of individual data tokens or constitutes a new kind of structure that expands the space of possibilities in the industry.

Semantic closure mechanism

During the fieldwork, we almost never saw raw data participating in organizational practices. The tokens are simply not practicable as such. As a whole, the data are voluminous and extremely detailed, suggesting that they could support a range of interpretations and insights. Yet, there is little actual information to work with in each individual data token, and turning their potential into facts about an audience is a far from trivial undertaking. Audiencemaking events that help to establish a new kind of audience product in the media market look quite different from the data tokens. For instance, an important event took place in August 2009, when a major industrial research firm confirmed some claims made by the company.

> Brands [advertisers] have been impressed with average campaign response rates of 25 percent. The richness of the interaction between Company's members and advertisers has also frequently been impressive. One example was a campaign organized by [Customer], which is a leading contact point for advice and guidance on bullying. The campaign was created to engage with 16- to 19-year-olds on this sensitive issue. Thirty-six percent of targeted members responded to the initial SMS [text message], and several of the responses revealed sensitive personal experiences and emotions. This type of engagement has convinced advertisers that mobile is a viable engagement medium for their target audiences.

(Industrial analyst report, August 2009)

The event is notable in that an external agency supports the claims about the new kind of audience by circulating them through its report. The document specifically reiterates metrics that define the audience members by their behaviour. While the company had already put forward such claims on numerous other occasions, the analyst report effectively frames them as factual statements by a seemingly independent actor. Other similar behavioural constructions of the audience are found throughout the empirical evidence. For instance, the manager for advertising operations (BMA) described the product in his interview as follows:

> BMA: Our [advertising] format is really good. It needs to be fine-tuned, but in general it is good: the response rate and all the behaviour we can generate – web traffic increases, coupon redeems and ROI [return on investment] for which it indeed culminates.

(Interview of Business Manager, Advertising (BMA) on 13 May 2009)

What makes it possible to conceive the audience as an interactive entity in the way that BMA does? The interactive characteristic contrasts with more traditional media. The construction of TV and radio audiences has historically revolved around the reception of media content by prescribed demographic segments, whereas the manager describes the new audience product as triggering and measuring behaviour. The shift from demographic to behavioural definition makes sense against the backdrop of the vastly improved measurability of behaviour. The essence of the new audience is not who it is but what it does. For instance, the rate at which the audience responds to advertising messages provides a good example of behavioural measures. It is referred to as the 'response rate' in the excerpt above, and, looking across our empirical evidence, the rate is one of the most important metrics the company uses to describe its audience.

The construction of the response rate metric presupposes suitable data and the means by which the data are combined together. Represented as a single number or a graph, the rate becomes part of the cognitive context for decision making and practical action. A concrete number can be pointed at, discussed and connected with many other events and measures, unlike an amorphous mass of CDRs. However, actual response rate readings could not form a foundation for other activities unless the mechanism by which they are produced remains stable over time. The rates are calculated by an algorithm that is embedded into the company's systems, filtering and combining data tokens according to a rigid procedure. The data are not coupled to a specific idea such as the response rate or any other metric that is brought into existence by programmatic operations. We observed a host of other metrics, including the number of active audience members, delivery of advertising messages and hyperlink clicks. These organizational metrics help to stabilize the focus on the inherently ambiguous audience. They render the audience product by producing its proportions on the specific dimensions of interest.

The data tokens are highly granular. They also capture a whole range of irrelevant, ambiguous and unexpected behavioural detail. For example, it cannot be decided, on the basis of data alone, if a repeated answer by the same member to an advertisement should be counted as one or two answers; or, what to do with a response to an advertisement that does not solicit any interaction. Such issues are not insignificant details. They indicate an important difference between a metric and the applications used to observe its actual readings. The response rate needs to be exactly the same irrespective of the application used to check its reading, which means that the metric cannot be solely an artefact of the software application and its user interface. The actual readings are expected to change constantly (though not too much) in order to be perceived as a reliable reflection of behavioural patterns outside the system, but this needs to happen in the context of steadfastly coded procedures.

By ‘semantic closure’ we mean a stable way to interpret the data for a specific purpose, which is embedded and stabilized in technology. It then becomes taken for granted by relevant stakeholders. The automatic and continuous calculation of response rates is an example of a mechanism that provides a semantic closure on the data. The metrics become (and must be) black boxes for organizational practices. They hide their internal complexity, provide continuously updated readings, and remain stable over time. The metrics express these features consistently in all of their implementations. By stabilizing a specific procedure for interpreting data, the response rate algorithm allows a massive reduction of potential readings, collapsing them into one that becomes actual. It turns all but meaningless data into specific information about the audience.

Pattern-finding mechanism

The employees observe the metrics using a variety of reporting software applications. However, the applications do more than just generate the semantic closures that maintain the metrics. They are tools that allow user intervention by setting the parameters on how data is filtered, combined and represented in the context of organizational practices. Using the applications, the employees can mine the data for various kinds of patterns beyond the few stable metrics. Let us start from an event in which a certain aspect of the audience became suddenly unavailable. The following excerpt depicts a situation in which a reporting system was perceived to fail in turning available data into information about the audience.

> X1 comes over [to our table] and asks how should the large-scale operation on the member base be targeted. MCM and BMMA point out that the operation should be started immediately, because next week it might be too late. [...] X1 asks, which members are to be terminated. [...] MCM ponders what is reasonable and what is not. He points to the coffee table discussion in which it had been decided that the Member experience reporting tool will not be [immediately] updated. Resulting from this, we now lack adequate information for the decision.
>
> (Observation log, 18 February 2009)

An outdated reporting application would hardly feel a problem if the data it represents do not matter. More specifically, the missing information appears against MCM's valid expectation of being able to elicit certain information from the data, which is based on his previous experiences on working with the tool. All in all, we identified five applications for retrieving, analysing and representing data on audience members (see Table 2). The applications enable employees to routinely represent aspects of the audience and its members, single out issues, and plan and execute both regular and *ad hoc* interventions. Many of the tools are used on a daily basis.

<Insert Table 2 here>

In contrast to the essentially rigid metrics, the logic of reporting applications is to enable multiple ways to arrange and summarize the voluminous data. The reporting applications are, first and foremost, user interfaces for querying multidimensional data. They enable employees to filter, combine and juxtapose data tokens, and to represent the results in tabular and visual forms. These representations often encapsulate organizational metrics discussed in the previous section. For instance, it is possible to compare the response rates for different advertisements in different geographical regions, between genders, and over time. The reporting applications help to uncover many patterns that may or may not be relevant, yet it is the data that ultimately set the boundaries and the possible paths for such explorations. The more data and dimensions a particular source offers, the more information a reporting application working with it can potentially reveal. The tools allow the situated judgement and inventiveness of employees to discover new avenues for making sense of the audience.

The pattern-finding mechanism is characterized by the role played by human operators, who need to devise strategies that could reveal more information from the data. Pattern-finding activities vary from mostly routinized activities to highly explorative attempts. In fact, we observed events that seem to express a different

form of pattern-finding mechanism in operation. These events are associated with manually crafted analyses based on custom database queries and using statistical packages to analyse the output. Apparent problems in the network infrastructure, inexplicable member behaviour, or the needs of business development could motivate such a novel cut into the data. Also, potential information in the data simply drew interest from some employees, who had consequently developed a habit of making casual data-mining exercises. The employees perceived and acted on the assumption that there is more information in the data than that which is being actualized by the current metrics and reporting applications.

Such exploratory opportunities are also harnessed by business development activities. Instead of precarious guesses about member behaviour and reactions to planned operations, it is sometimes possible to test assumptions by using reporting applications or by crafting a custom analysis. For instance, on one occasion it was necessary to dig deeper into the nature of member engagement with the advertisements. MCM, who was responsible for the member analytics, suggested studying the matter from the data. In a matter of hours he put together a graph depicting the speed of responses of different demographic groups. The visualization revealed interesting patterns beyond the aggregate response rate. For instance, it was found that the members either answer within a few minutes of the arrival of a message or are unlikely to engage the advertisement at all. Proposing such an analysis would have made little sense without the readily available data. The data pool provides a kind of laboratory environment where emerging ideas can be tested.

Learning from custom analyses also feeds back to the further development of measurement arrangements. Free-form explorations into the data can serve as initial steps for the development of new metrics and reporting applications. To summarize, the pattern-finding mechanism is made possible and boosted by the highly granular and comprehensive data generated by the digital network infrastructure. It also points to an interesting feature of the space of possibilities that the data open up. It is taken for granted that there is potential information in

the pool of data, but the amount of that potential information is unknown. The boundaries of pattern-finding are therefore *a priori* undefined, for it is not known in advance what can be done with the data.

The employees can query, tabulate and visualize patterns in the data using the reporting applications. The applications allow the activation of a pattern-finding mechanism. On the one hand, pattern-finding also provides a semantic closure on data tokens, but, on the other hand, the activation of the pattern-finding mechanism involves trying out and choosing between different semantic closures, not just reading a prescribed metric. Both the actual patterns and the ways to compile them can change, and, unlike the semantic closure mechanism, stability is not an overarching concern. The mechanism modulates between furthering established paths of semantic closure and the establishment of new ways to make sense of the data. The metrics and the use of reporting applications are the foundation for numerous reporting practices at the office.

Framing mechanism

The most generic reporting practice at the company is a weekly office meeting in which senior managers give brief updates on different aspects of the business to the staff. The meetings are held in the office lobby area as standing sessions without a formal decision-making function. For instance, we observed an event in which a senior manager (X3) asks about the size of the member base and tells briefly about the status of advertising sales.

> X3 asks about the number of members. MCM answers that we have 75000 primary SIM card holders. X3 says that the number of top-ups is above the budgeted and advertising sales are proceeding fairly well, even though achieving the budgeted sales will require very hard work. He continues to point out that the revenues of biggest media companies have dropped thirty per cent meaning that the market is really in a recession.
>
> (Observation log, 10:00 on 9 March 2009)

On an occasion such as the office meeting, the construction of an advertising audience becomes a largely interpretive exercise. The discussion about the overall audience size offers a good example. It may seem a simple, unambiguous number. MCM chooses to answer in terms of subscribers who use the company SIM card as their primary mobile phone subscription. This implies that that there are also other ways to count the number of members. For instance, the count would be different if it were reported as the number of people who hold a company SIM card. In a similar manner, the fact that sales are lagging behind targets is framed by the senior manager as fairly good by contrasting it to the current market conditions. The selection, timing and presentation of facts can matter just as much as information from the data. The office meeting was usually re-interpreted over lunch. In the lunch discussions, employees' views ranged from suggesting slightly different twists to the reported matters to debating what was the message that senior managers truly conveyed.

> People discuss some work-related matters over lunch. UED ponders that the tone in the office meeting was moderately positive. Others agree. HT jokes about running away to Bahamas with investors' money; AA continues that we are merely producing reports. Let's leave somebody behind to keep churning out the reports.
>
> (Observation log, 12:56 on 9 March 2009)

The comment about reports by AA is particularly revealing due to its inherent sarcasm. He acknowledges the importance of reports and reporting activities yet describes them as framing – 'we are merely producing reports'. AA thus suggests that reporting itself has become the focus of their work, not the things that are being reported. By framing, therefore, we mean the way in which the metrics and patterns observed in the data are brought to bear upon daily operations. The above comment is sarcastic because the employees are well aware that the mere practice of reporting is not enough. Behind the oral accounts put forward by senior managers at the office meetings, there are numerous reporting practices carried out in daily,

weekly and monthly cycles in the organization. In the context of such practices, employees selectively associate metrics and patterns found in the data with other sources of information, trends and objectives. The following interview excerpt shows how this occasionally went too far, generating reports which were too complex and which then required re-framing to again be useful.

> HBD: X2 had one chap [in the local sales office] who compiled the statistics. And Operations team aggregated some other numbers and from these it was put together. [...] I was perhaps sometimes a little bit sceptical. We had sort of papers that incorporated 20 KPIs [key performance indicators]. For all those I told X3 and CEO that this is too complex. [...] In fact, I kept simplifying those numbers into Excel for myself even after we had the more sophisticated reporting, so that I could do the follow up [on member acquisition] compared to the earlier period.
>
> (Interview with Head of Brand and Design (HBD) team on 16 September 2009)

Manually compiled PowerPoint presentations and Excel spreadsheets have a specific advantage over the pre-compiled metrics and the reporting applications. People are able to select readings from different sources, combining and juxtaposing them with different tactics. In doing so, it is possible to strategically guide the interpretation of information to address issues from a specific perspective. There was often a lot of discussion on what a specific metric means for the task in hand, or what readings should be shown on a particular occasion or for specific material. For instance, it was not always clear how to count the number of active audience members against those lying dormant in the database. While this allows discretion and a degree of strategic ambiguity, without the data, metrics and reporting applications no credible reporting about the audience would have been possible.

In the three events described above, we perceive a mechanism that frames facts emerging from the data pool by virtue of the semantic closure and pattern-finding mechanisms. The purpose of the practical framing of facts is to more easily evoke

certain interpretations while shunning others. At the same time, it produces new meaning that can be grasped only when the relationships between heterogeneous pieces of information are considered. Without such framing, the risk is that produced facts do not stand out or, even worse, are placed against an unfavourable background from the perspective of the company or an individual employee. The data pool alone is not enough to account for such a generic framing mechanism, which is activated, rather, at the encounter of interpretive agency and forms of aggregate data. The framing mechanism would merely produce an empty frame without the metrics, tabulations and data visualisations generated by the semantic closure and the pattern-finding mechanisms.

DISCUSSION

The new audience product is defined and maintained by the operation of semantic closure, pattern-finding and framing mechanisms that operate on the raw CDR data. The three mechanisms are nested so that an output from one feeds the other (see Appendix 1). This allows information about the audience to cascade through metrics, reporting applications and practices, becoming richer and more relevant for audiencemaking practices at every step. Table 3 summarizes the type of activating condition, observable entities and the typical operation of each mechanism.

<Insert Table 3 here>

Media companies have traditionally sold advertising space on the basis of the predicted amount of attention that a particular placement will attract, while the effective audience (those who actually saw the advertisement) used to be inferred *post hoc* from a sample of consumers participating in industrial audience measurement panels (Napoli, 2003). Our case study confirms and deepens the insight that the "institutionally effective audience" (Ettema and Whitney, 1994) is not made of people but data. What cannot be measured cannot be verified to the advertisers and thereby cannot be part of the audience product. Against this background, the data generated by the digital network infrastructure introduces a

major shift (Bermejo, 2009; Carr, 2008). The nexus of value creation shifts from obtaining valid and reliable samples of people's media consumption to analysing the audience from the extant data. Observing mobile phone users on the street would not help the company understand the audience because, paradoxical though this statement may seem, the audience is not out there but constructed from the data. In the following section, we elaborate the findings of retroductive analysis by theorizing a more generic mechanism and by identifying properties of the data pool. Finally, we will discuss the validity of the findings.

Information actualization

The advertising-funded telecommunications operator is, in certain respects, a relatively straightforward venture. The data pool offers a space of possibilities for the company to create a new kind of advertising platform with which to compete against both traditional advertising businesses and subscription-based network operators. A key assumption underpinning the venture is that the CDRs contain an informative potential, that can be extracted through automatic and manual elaborations, and then used to fuel audiencemaking operations. However, it is important to understand that valuable information is only potential in the data. It is something that can become expressed through certain events, or not. The data pool contains differences that are not *prima facie* meaningful (Bateson, 2000). We have shown in the analysis how, under certain conditions, these differences have an effect in the audiencemaking events (Bateson, 2000, p. 459; Kallinikos, 2006, p. 60–61; McKinney and Yoos, 2010). The relationship between the data as raw material and the audience as a product can be understood through the Aristotelian dichotomy of potentiality versus actuality (Cohen, 2012).

Let us rely on a generally accepted understanding of actuality as the fulfilment of a potentiality, while potentiality indicates the possibility for something to happen, or come into being. The actual and potential are defined in relation to each other, one complementing the other. Aristotle argues in the *Metaphysics* that actuality stands to potentiality "as that which has been shaped out of some matter is to the matter from

which it has been shaped" (1048b1-3 as in Cohen, 2012). Here, if we understand the data as the digital matter from which information is extracted, the three mechanisms constitute a set of *information actualization mechanisms*. Information actualization describes various ways to exploit the new space of possibilities that exists by virtue of pooling vast amounts of digital data.

The idea of information as actualized potential is analogous to the classic marble statue example. Russell (1994, p. 180) writes "'a block of marble is a potential statue' means 'from a block of marble, by suitable acts, a statue is produced.'" The block of marble (data) neither determines the existence of the statue nor its shape (information), but it is equally true that the statue could not appear out of nothing. The potential does not exist in material alone, but requires the availability of means to transform the material into something else. It takes a combination of suitable skills, actions and material for something to happen or come into being.

Properties of the data pool structure

The foundations of the semantic closure and pattern-finding mechanisms we have identified lie in the structural properties of the data pool. The practical conditions for their emergence stand in the sheer amount of data and the technological capacity to simultaneously filter and combine a large number of tokens. We identify three properties that define the data pool structure: the *comprehensive*, *granular* and *unbounded* characteristics of the data pool.

To begin with, the digital data tokens matter because the digital network infrastructure automates much of the data collection. In traditional media, this is done by separate measurement devices distributed to a small subset of consumers. The collected data is then limited to carefully planned samples geared to predefined purposes, whereas in the present digital ecosystem the behaviour of the whole user base is captured implicitly by the infrastructure. There is no need to distribute and maintain the expensive metering devices. Importantly, the massive amount of data generated by the digital infrastructure is not a sample but the census of the activity

in the network. The data pool can be said to be a *comprehensive* collection of user behaviours.

The digital network infrastructure not only automates the data collection, but also generates records which are qualitatively different, as compared to earlier audience measurement arrangements. CDRs were not designed for audiencemaking purposes. They dissolve media use into discrete clicks and messages. It is from the pool of such extremely *granular* behavioural traces that meaningful behavioural patterns have to be reassembled by recourse to analytic operations (Kallinikos, Aaltonen and Marton, 2013). If the data collection was earlier framed as surveying predefined consumer segments and categories, those have to be now produced *a posteriori* from the extant data. The meaning lost in the extreme granularity of the data is, however, compensated by the vastly expanded opportunities to aggregate, align and juxtapose digital data tokens against each other (Kallinikos, 2006; Kallinikos, Aaltonen and Marton, 2013).

Finally, the individual data tokens represent ephemeral behavioural episodes, which give them a "use-agnostic" character (Kallinikos, 2012). The data are loosely coupled with the uses to which they are actually put and may not immediately seem able to answer any relevant question. They exist as an open-ended potential, to be explored in a variety of ways and to different ends. Importantly, the pool of agnostic data tokens leaves the boundaries of such explorations open and undefined. This makes the space of possibilities emerging from such data look characteristically *unbounded*. What can be done with the data depends on the availability and activation of specific information actualizations mechanisms.

Table 4 summarizes the three properties of digital data in the case. The properties are hardly idiosyncratic to the case, but we acknowledge that other cases may also exhibit other properties (Ekbia, 2009; Faulkner and Runde, 2010; Kallinikos, Aaltonen and Marton, 2013; Yoo, Henfridsson and Lyytinen, 2010). While comprehensiveness and unboundedness are attributable only to the data pool as a

whole, granularity could be understood as a property of the individual data token object. The former two are thus emergent properties (Elder-Vass, 2005, 2007); they appear as large amounts of data tokens and are managed in relation to each other. The presence of emergent properties suggests that the data pool is a new kind of structure and should not be considered just a heap of data. It has causal powers that support the activation of the mechanisms we have found through the analysis of empirical evidence.

<Insert Table 4 here>

Let us briefly qualify the three properties and explain why we think they are either emergent or resultant properties (Elder-Vass, 2007). To begin with, comprehensiveness cannot obviously be attributed to an individual data token. It results from the collection of the totality of behavioural events in the network and, unlike a sample, allows individual interaction with each member. The case is different with regard to granularity, which, in our case, concerns the resolution at which people's media use is recorded. A data token represents a single member interaction and, in this respect, granularity is a resultant attribute of individual objects in the data pool. Nevertheless, a highly granular pool of data tokens enables the data to be explored by many more combinations than a less granular pool of data would allow. The third property, unboundedness, and the other two properties above, are interrelated. The potential of the data to inform about many unforeseen issues would be limited without the comprehensiveness and granularity of the data. It is the combination of breadth (comprehensiveness) and resolution (granularity) that explode the number of potential questions that can be asked from the data. Unboundedness is thus an emergent property.

The validity of the findings

The three mechanisms described in this study are candidates for causal explanations of the observed events. The critical realist metatheory requires the results to be presented so that they can be tested against alternative hypotheses,

and it has been argued that studies should include an assessment of the identified mechanisms against other possible explanations (Bygstad, 2010; Runde, 1998; Wynn and Williams, 2012). We first consider an alternative kind of explanation to the audiencemaking events and then discuss the analysis against a set of evaluation criteria for causal explanations.

A possible alternative explanation could be based on the assumption that the properties of digital data have no significant impact on audiencemaking events and, consequently, on the audience sold by the company. One could try to argue that it is possible to understand the audience in terms of the coalescing of interpretive acts. The response rate and other characteristics of the audience product could be analysed as choices made by the actors and not as outcomes shaped by the mechanisms that emerge from the digital data. The alternative explanation would then centre on negotiations and interpretations in the process of constructing the audience. However, important aspects of the case escape this kind of explanation. The audience members are found to behave in unexpected ways in the data; they surprise employees and shape their plans and expectations. Furthermore, the occasional inability to turn data into information would not hinder action if the data pool were not making a difference to organizational practices. The alternative explanation limited to the interpretive dimension of organizational practices would fail to recognize the specific ways in which the data enabled and constrained the construction of the audience.

Runde (1998) proposes four principles for evaluating a retroductive causal explanation. A causal hypothesis is considered plausible and well-formed if the candidate mechanism: is taking part in the situation where the observed consequence occurred; is a plausible cause of an event that needs an explanation; is deemed sufficient to cause the aspect of the event under scrutiny; expresses a degree of causal depth (it has explanatory power). In regard to the first principle, the three structural properties of the data pool and the three mechanisms are clearly implicated in audiencemaking events. Second, the reactions and

interpretations with respect to the data are events that warrant an explanation, since they are critical to the success of the company. We have shown how important aspects of the events could not be understood without unpacking the role that the data pool plays in their unfolding. Third, our explanation is sufficient in that we retroduced a set of related mechanisms that, if they were real, would explain why the observed events construct the audience in the way they did. We aimed to postulate only the structures, mechanisms and powers that it is necessary to take into account at the level of abstraction at which we are developing our argument. The explanation does not exclude other intervening or countervailing causal powers. For instance, we have identified the presence of an interpretive element contributing to the framing mechanisms that is involved in constructing the new kind of audience. Fourth, the argument has causal depth. It explains how an advertising audience is constructed in the digital ecosystem by reference to specific mechanisms and the data pool structure.

CONCLUSIONS

In this article, we have demonstrated the use of critical realism for studying the production of data-driven products and services. The argument was substantiated by analysing how a telecommunications operator transforms agnostic data from a network infrastructure into valuable information about a new kind of advertising audience. Critical realism helped to pin down audiencemaking events against a relevant industrial background and then analyse how the audience is manufactured from the data. The findings are based on a single case study, but our contribution toward understanding the mechanisms of information actualization could be broadly validated.

The findings are relevant and timely. Information systems do not just store, process and transfer data, but they also generate vast amounts of new data. New data may have initially been generated for only peripheral uses (such as maintaining the network itself), but they are also increasingly recognized as raw material for new products and services. Indeed, products such as advertising audiences, securities,

insurances and many kinds of ratings could be called ‘data-based’ rather than data-driven, for they are made out of data (Redman, 2008). Recently, there has been a lot of excitement and discussion about the opportunities of ‘big’ and ‘open’ data. In several ways, the research site represents many of those organizations that execute novel business models around what is perhaps vaguely termed Big Data (Boyd and Crawford, 2012).

Whether data-based business opportunities can be realized depends on an organizational capability to harness the potential embedded in newly available digital data. Many organizations are at a loss with these opportunities. They either sit unknowingly on top of an enormous resource or lose themselves in the morass of meaningless analytics (Aaltonen, 2012; Day, 2003). Building metrics and developing reporting tools and practices are seldom perceived as the most interesting activities in an office, but understanding them is critically important to an increasing number of businesses. The data has no value without the arrangements that can realize its potential; our study is a concrete example how those arrangements can be studied and offers a set of mechanisms as a starting point.

More generally, our study differs from the body of IS literature in which computing is “conceptualized as a discrete symbolic representations of something in the *real* world” (Yoo 2010, p. 218). The individual data tokens may be understood to represent actions of flesh-and-blood human beings, but the audience does not have such a clear, external referent. The aggregate of digital data (what we define as the data pool) is real matter with emergent properties. The product is literally manufactured from such raw digital material. Supported by a critical realist metatheory, IS scholars can be at the forefront of explaining the transition from the mere processing (or *reading* as in Zuboff, 1988; Kallinikos, 1999) of technological representations to new socio-technical configurations that involve the construction of new products and forms of value creation on digital data. Wikipedia and open source software development are good examples (Aaltonen and Kallinikos, 2013; Benkler, 2006), but there are many others.

We believe that digital materiality needs to be studied intensively, that is, by theorizing emergent properties specific to the digital ecosystem. While we are sympathetic to the agenda set forth by Leonardi (2010), the analysis of digital materiality as emergent properties and mechanisms raises issues with respect to the definition of materiality as "practical instantiation of theoretical ideas" and "what is significant in the explanation of a given context" (Leonardi, 2010). These two definitions provide useful perspectives, but they exclude certain aspects regarding how the digital ecosystem matters in business. Digital data, in the form of structures such as a data pool, do more than just instantiate theoretical ideas. Ideas often require material underpinnings to be conceivable in practical terms. There is no reason why ideas should pre-exist materiality – some may, but the opposite situation can also exist. Working hands-on with materials stimulates curiosity and imagination, making it possible to develop new ideas (Dourish, 2001). We have shown throughout our study that a data pool defines a space of possibilities. It is the matter within which a number of work efforts are imagined, conceived and executed. Our theorizing generally agrees with Leonardi's second definition, but it is important to point out that the emergent properties of digital data are not straightforwardly read off from empirical observations. Understanding 'material' as that which matters for a given activity is a good starting point (cf. Latour, 1999). However, we also need robust conceptual tools to analyse how generic attributes of the digital ecosystem matter in specific industries and organizational settings.

References

Aaltonen, A. (2012) The Beauty and Perils of Metrics, *Mercury Magazine* 1(3): 56-59.

Aaltonen, A. and Kallinikos, J. (2013) Coordination and learning in Wikipedia: Revisiting the dynamics of exploitation and exploration, *Research in the Sociology of Organizations* 37: 161-192.

Archer, M. S. (1982) Morphogenesis Versus Structuration: On Combining Structure and Action, *The British Journal of Sociology* 33(4): 455-483.

Archer, M. S. (1998) Introduction: Realism in the Social Sciences, in M. Archer, R. Bhaskar, A. Collier, T. Lawson and A. Norrie, (eds.) *Critical Realism: Essential Readings*, New York: Routledge, pp. 189-205.

Baskerville, R. L. and Myers, M. D. (2002) Information Systems as a Reference Discipline, *MIS Quarterly* 26(1): 1-14.

Bateson, G. (2000) *Steps to an Ecology of Mind*, Chicago: Chicago University Press.

Becker, H. S. (2007) *Writing for Social Scientists: How to Start and Finish Your Thesis, Book, or Article*, Chicago: University of Chicago Press.

Benbasat, I. and Zmud, R. W. (2003) The Identity Crisis within the IS Discipline: Defining and Communicating the Discipline's Core Properties, *MIS Quarterly* 27(2): 183-194.

Benkler, Y. (2006) *The Wealth of Networks: How Social Production Transforms Markets and Freedom*, New Haven: Yale University Press.

Bermejo, F. (2009) Audience Manufacture in Historical Perspective: From Broadcasting to Google, *New Media & Society* 11(1/2): 133-154.

Bhaskar, R. (1998) Philosophy and Scientific Realism, in M. Archer, R. Bhaskar, A. Collier, T. Lawson and A. Norrie, (eds.) *Critical Realism: Essential Readings*, New York: Routledge.

Bhaskar, R. (2008) *A Realist Theory of Science*, New York: Routledge.

Boyd, D. and Crawford, K. (2012) Critical Questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662-679.

Bratich, J. Z. (2005) Amassing the Multitude: Revisiting Early Audience Studies, *Communication Theory* 15(3): 242-265.

Brewer, J. D. (2000) *Ethnography*, Buckingham: Open University Press.

Bygstad, B. (2010) Generative Mechanisms for Innovation in Information Infrastructures, *Information and Organization* 20(3-4): 156-168.

Carr, N. G. (2008) *The Big Switch*, New York: W. W. Norton & Company.

Cohen, S. Marc (2012) Aristotle's Metaphysics, *The Stanford Encyclopedia of Philosophy (Summer 2012 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2012/entries/aristotle-metaphysics/>

Crotty, M. (1998) *The Foundations of Social Research: Meaning and Perspective in the Research Process*. London: SAGE Publications.

Day, G. S. (2003) Creating a Superior Customer-relating Capability, *MIT Sloan Management Review* 44(3): 77-82.

Dourish, P. (2001) *Where the Action Is*, London: The MIT Press.

Easton, G. (2010) Critical realism in case study research, *Industrial Marketing Management* 39: 118-128.

Ekbia, H. R. (2009) Digital Artifacts as Quasi-Objects: Qualification, Mediation, and Materiality*, Journal of the American Society for Information Science and Technology* 60(12): 2554-2566.

Elder-Vass, D. (2005) Emergence and the Realist Account of Cause, *Journal of Critical Realism* 4(2): 315-338.

Elder-Vass, D. (2007) For Emergence: Refining Archer's Account of Social Structure, *Journal for the Theory of Social Behaviour* 37(1): 25-44.

Ettema, J. S. and Whitney, D. C. (1994) The Money Arrow: An Introduction to Audiencemaking, in Audiencemaking: How the Media Create the Audience, in J. S. Ettema and D. C. Whitney, (eds.) *Sage Annual Reviews of Communication Research*, London: SAGE Publications, pp. 1-18.

Faulkner, P. and Runde, J. (2009) On the Identity of Technological Objects and User Innovations in Function, *Academy of Management Review* 34(3): 442-462.

Faulkner, P. and Runde, J. (2010) The Social, the Material, and the Ontology of Non-Material Objects, in Judge Us Seminar (University of Cambridge, UK, 2010).

Faulkner, P. and Runde, J. (2013) Technological Objects, Social Positions, and the Transformational Model of Social Activity, *MIS Quarterly* 37(3): 803-818.

Flick, U. (2004) Triangulation in Qualitative Research, in U. Flick, E. von Kardoff and I. Steinke, (eds.) *A Companion to Qualitative Research,* London: SAGE Publications, pp. 178-183.

Kallinikos, J. (1999) Computer-Based Technology and the Constitution of Work: A Study on the Cognitive Foundations of Work, *Accounting, Management & Information Technology* 9(4): 261-291.

Kallinikos, J. (2006) *The Consequences of Information: Institutional Implications of Technological Change*, Northampton, MA: Edward Elgar Publishing.

Kallinikos, J. (2012) The Allure of Big Data, *ParisTech REVIEW* 16.11.2012 [online] http://www.paristechreview.com/2012/11/16/allure-big-data

Kallinikos, J., Aaltonen, A. and Marton, A. (2013) The Ambivalent Ontology of Digital Artifacts, *MIS Quarterly* 37(2): 357-370.

Langley, A. (1999) Strategies for theorizing from process data, *Academy of Management Review* 24(4): 691-710.

Latour, B. (1999) *Pandora's Hope: Essays on the Reality of Science Studies*, Cambridge, MA: Harvard University Press.

Lee, A. S. (2010) Retrospect and Prospect: Information Systems Research in the Last and Next Twenty-Five Years, *Journal of Information Technology* 25(4): 336–348.

Leonardi, P. M. (2010) Digital Materiality? How Artifacts without Matter, Matter, *First Monday* (15:6), [online journal] http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3036/2567

Leonardi, P. M., Nardi, B. A. and Kallinikos, J. (2012) *Materiality and Organizing: Social Interaction in a Technological World.* Oxford, Oxford University Press.

Lyytinen, K. and Yoo, Y. (2002) The Next Wave of Nomadic Computing, *Information Systems Research* 13(4): 377-388.

McKinney, E. H. and Yoos, C. J. (2010) Information about Information: A Taxonomy of Views, *MIS Quarterly* 34(2): 329-344.

Mingers, J. (2004) Real-izing Information Systems: Critical realism as an Underpinning Philosophy for Information Systems, *Information and Organization* 14(2): 87-103.

Morley, D. (2006) Unanswered Questions in Audience Research, *The Communication Review* 9(2): 101-121.

Mutch, A. (2010) Technology, Organization and Structure – A Morphogenetic Approach, *Organization Science* 21(2): 507-520.

Napoli, P. M. (2003) *Audience Economics: Media Institutions and the Audience Marketplace*, New York: Columbia University Press.

Orlikowski, W. J. (2007) Sociomaterial Practices: Exploring Technology at Work, *Organization Studies* 28(9): 1435-1448.

Orlikowski, W. J. and Barley S. R. (2001) Technology and Institutions: What Can Research on Information Technology and Research on Organizations Learn from Each Other? *MIS Quarterly* 25(2): 145–165.

Redman, T. C. (2008) *Data Driven*, Boston: Harvard Business Press.

Runde, J. (1998) Assessing Causal Economic Explanations, *Oxford Economic Papers* 50(2):151-172.

Runde, J., Jones, M., Munir, K. and Nikolychuk, L. (2009) On Technological Objects and the Adoption of Technological Product Innovations: Rules, Routines and the Transition From Analogue Photography to Digital Imaging, *Cambridge Journal of Economics* 33(1): 1-24.

Russell, B. (1994) *History of Western Philosophy*, Routledge: London.

Sayer, A. R. (2000) *Realism and Social Science*, London: SAGE.

Sismondo, S. (1993) Some Social Constructions, *Social Studies of Science* 23(3): 515-553.

Simon, H. A. (1996) *The Sciences of the Artificial*, Cambridge, MA: The MIT Press.

Walsh, D. (1998) Doing Ethnography, in C. Seale, (ed.) *Researching society and culture*, London: SAGE Publications, pp. 217-232.

Weick, K. E. (1995) What Theory Is Not, Theorizing Is, *Administrative Science Quarterly* 40(3): 385-390

Wynn, D. and Williams, C. K. (2012) Principles for Conducting Critical Realist Case Study Research in Information Systems, *MIS Quarterly* 36(3): 787-810.

Yin, R. K. (2003) *Case Study Research: Design and Methods*, London: SAGE.

Yoo, Y. (2010) Computing in Everyday Life: A Call for Research on Experiential Computing, *MIS Quarterly* 34(2): 213-231.

Yoo, Y., Boland, R. J. Jr., Lyytinen, K. and Majchrzak, A. (2012) Organizing for Innovation in the Digitized World, *Organization Science* 23(5): 1398-1408.

Yoo, Y., Henfridsson, O. and Lyytinen, K. (2010) The New Organizing Logic of Digital Innovation: An Agenda for Information Systems Research, *Information Systems Research* 21(4): 724-735.

Zuboff, S. (1988) *In the Age of The Smart Machine: The Future of Work and Power*, New York: Basic Books.

Figure and table legends

Table 1

Type of Evidence	Quantity	Details
Observation log	62 days	13 February 2009 – 15 May 2009
Interviews (during the fieldwork period)	34	26 different informants
Press releases	26	November 2006 – May 2010
Blog posts (on the company website)	60	November 2006 – May 2010
Intranet usage statistics	335 days	July 2008 – May 2009
Documents	340	Reports, intranet pages, etc.
Instant messaging logs	59	Conversations between employees
Photographs	147	Meetings, office events, etc.
In-situ analysis		
Weekly summaries	14	One per observation week
Tailored interview guides	34	One per interview

Table 2

System	Data source	Purpose
Advertising reporting	Network infrastructure	Reporting on advertising delivery and member interactions with advertisements
Customer service system	Call centre	The management of customer service requests
Member experience reporting	Network infrastructure	The analysis of subscriber behaviour in the network
Web survey tool	Online forms	A tool for creating and reporting web surveys
Website traffic analysis	Network infrastructure	The analysis of company website traffic

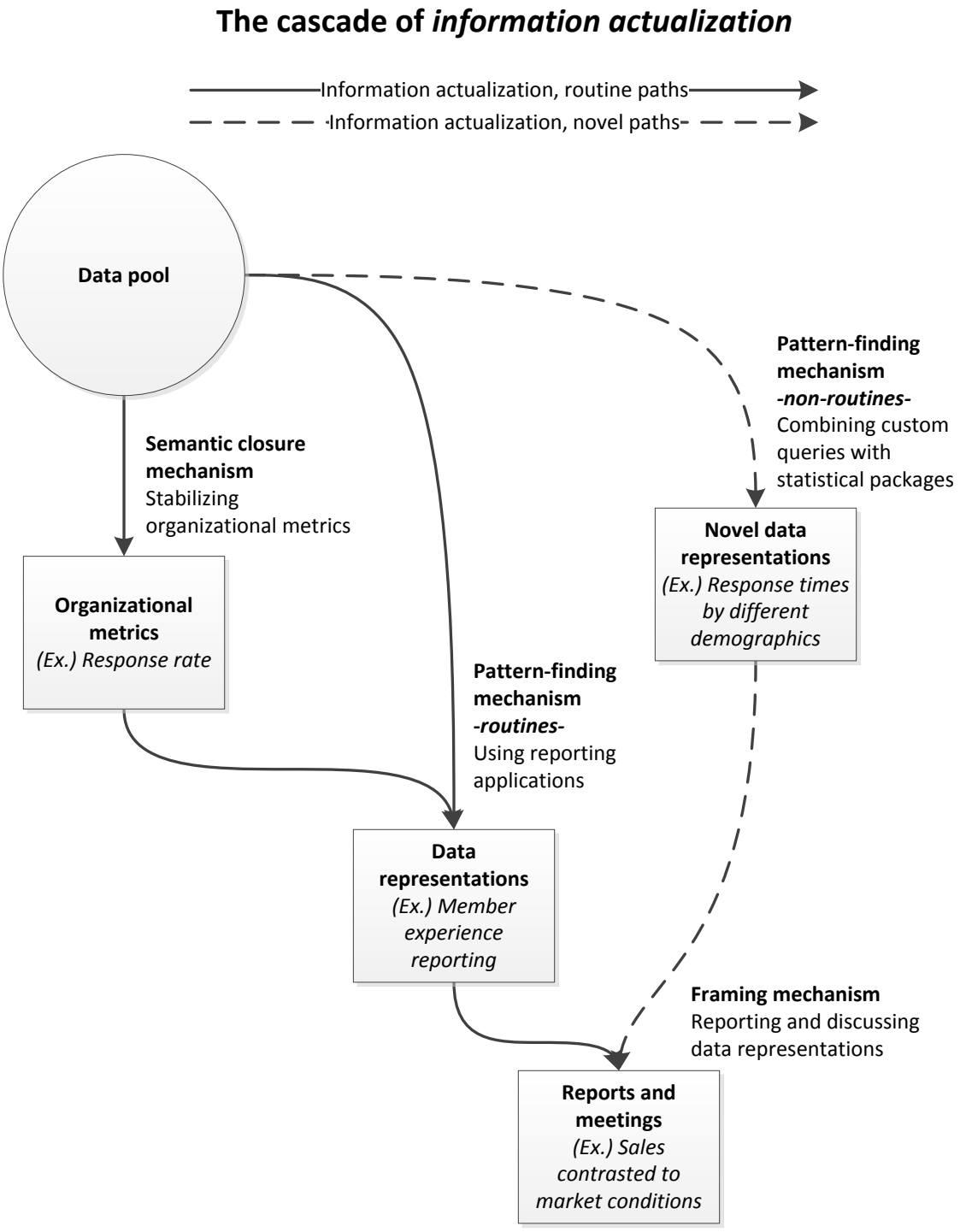
Table 3

Mechanism	Activating condition	Observable entities	Typical operation
Semantic closure	The execution of program code	Metrics	Through stabilization of a metric, a continuous change can be observed from a fixed viewpoint
Pattern-finding	The use of reporting applications; custom database queries combined with the use of statistical packages	Tabulated and visual representations of aggregate data	Trying out and choosing between different ways to look at the data enables eliciting informative patterns
Framing	Reporting practices	Presentations, spreadsheets, verbal accounts etc. that contain representations of aggregate data	The production of more information by connecting the data to other data sources with respect to a broader context

Table 4

Property	Type	Description
Comprehensive	Emergent	The data is the census of activity in the system (not a sample)
Granular	Resultant	The data tokens break a referent reality into meaningless behavioural episodes
Unbounded	Emergent	The boundaries of data-driven understanding are not known in advance

Appendix 1


The cascade of *information actualization*
Information actualization, routine paths
Information actualization, novel paths
Data pool
Semantic closure mechanism
Stabilizing organizational metrics
Organizational metrics
(Ex.) Response rate
Pattern-finding mechanism
-non-routines-
Combining custom queries with statistical packages
Novel data representations
(Ex.) Response times by different demographics
Pattern-finding mechanism
-routines-
Using reporting applications
Data representations
(Ex.) Member experience reporting
Framing mechanism
Reporting and discussing data representations
Reports and meetings
(Ex.) Sales contrasted to market conditions

Authors' bios

Aleksi Aaltonen

Aleksi Aaltonen holds a PhD in Information Systems from the London School of Economics and Political Science. He is broadly interested in technologies of organizing, and has written about business metrics, digital artifacts, data-driven practices in mobile advertising, and the governance of social production. His publications have appeared in outlets such as *MIS Quarterly* and *Research in the Sociology of Organizations*.

Niccolo Tempini

Niccolo Tempini is a PhD Candidate in Information Systems at the London School of Economics and Political Science, with a background in philosophy (BA, MA). His research focuses on organizations developing open and distributed networks for the generation, collection and analysis of big amounts of data. This research is relevant for the understanding of emerging organizational forms, work practices and mechanisms of information production and dissemination.

Endnotes

[i] By metatheory we refer to reasoning behind empirical research designs; a framework that provides the rationale and practical guidance on how the different aspects of research are brought together into a coherent argument. The term is largely synonymous with theoretical perspective (Crotty 1998), yet 'metatheory' communicates explicitly the idea of theory about research and distinguishes it, in our case, from substantive theorizing of technology in particular settings.

[ii] A mobile virtual network operator (MVNO) is a telecommunications operator that does not own a physical network infrastructure but leases it from another operator.

[iii] We are not allowed to reproduce an actual CDR from the research site.

[iv] Advenage SMS Gateway Router 1.0 documentation